# German credit data

## # 과제

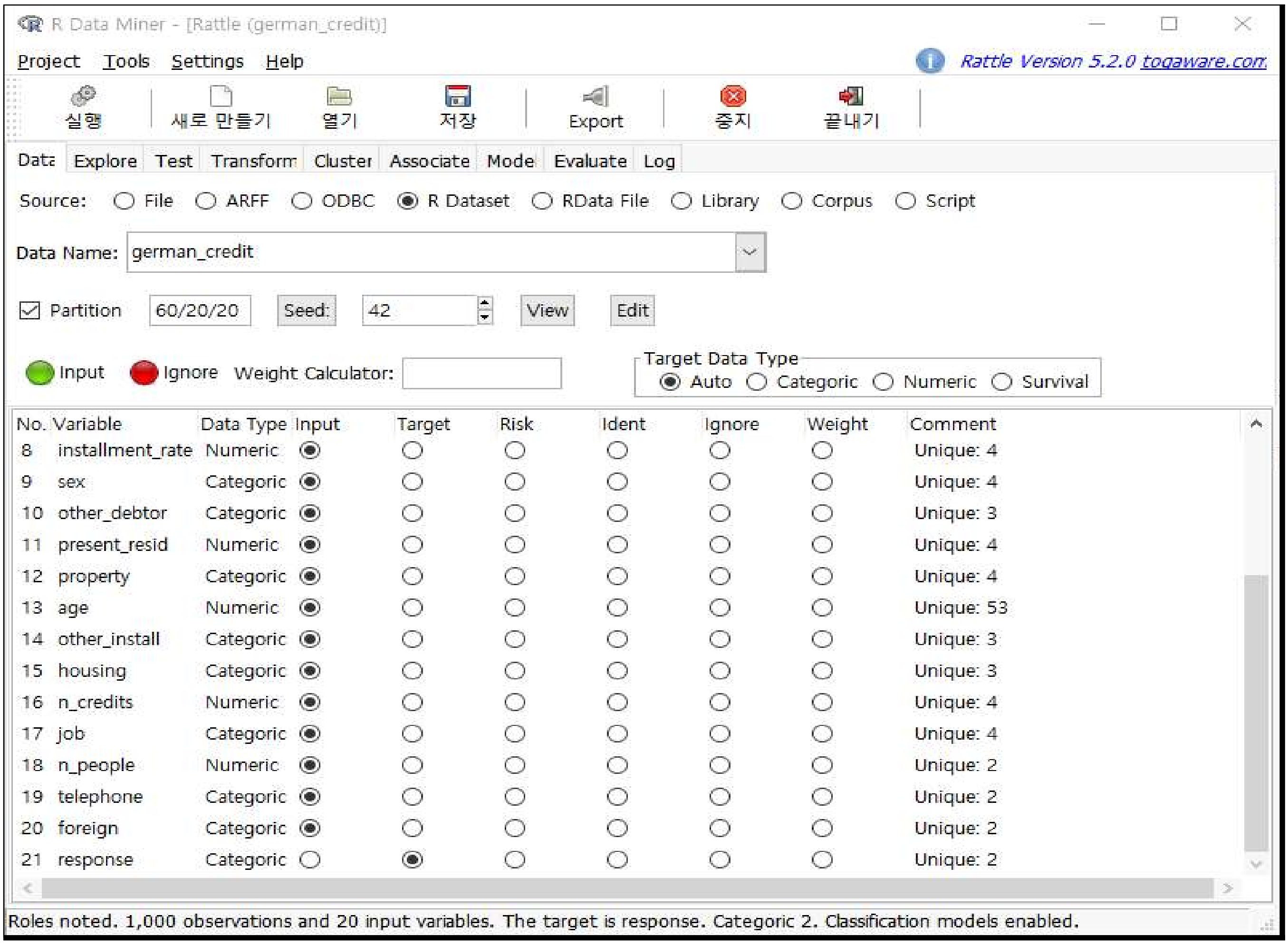
▶ German credit data를 train/validation/test의 비율을 60/20/20의 비율로 나누고, 로지스틱 회귀와 Decision Tree 모형을 적합하고, test 데이터의 오분류표를 구하라.

▶ FP와 FN의 loss를 1:8로 지정하여 로지스틱 회귀와 Decision Tree 모형을 적합하고, test 데이터의 오분류표를 구하라.

## # 코드

```
# Data Loading
german_credit <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data")

# Variable Naming
colnames(german_credit) <- c("chk_acct", "duration", "credit_his", "purpose",
                             "amount", "saving_acct", "present_emp", "installment_rate", "sex", "other_debtor",
                             "present_resid", "property", "age", "other_install", "housing", "n_credits",
                             "job", "n_people", "telephone", "foreign", "response")

# Response variable setting
german_credit$response <- german_credit$response - 1
german_credit$response <- as.factor(german_credit$response)

# Data Summary
str(german_credit)

# Rattle
library(rattle)
rattle()
```

## # Decision Tree


R Data Miner - [Rattle (german_credit)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
실행 새로 만들기 열기 저장 Export 중지 끝내기
Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ○ File ○ ARFF ○ ODBC ● R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Data Name: german_credit

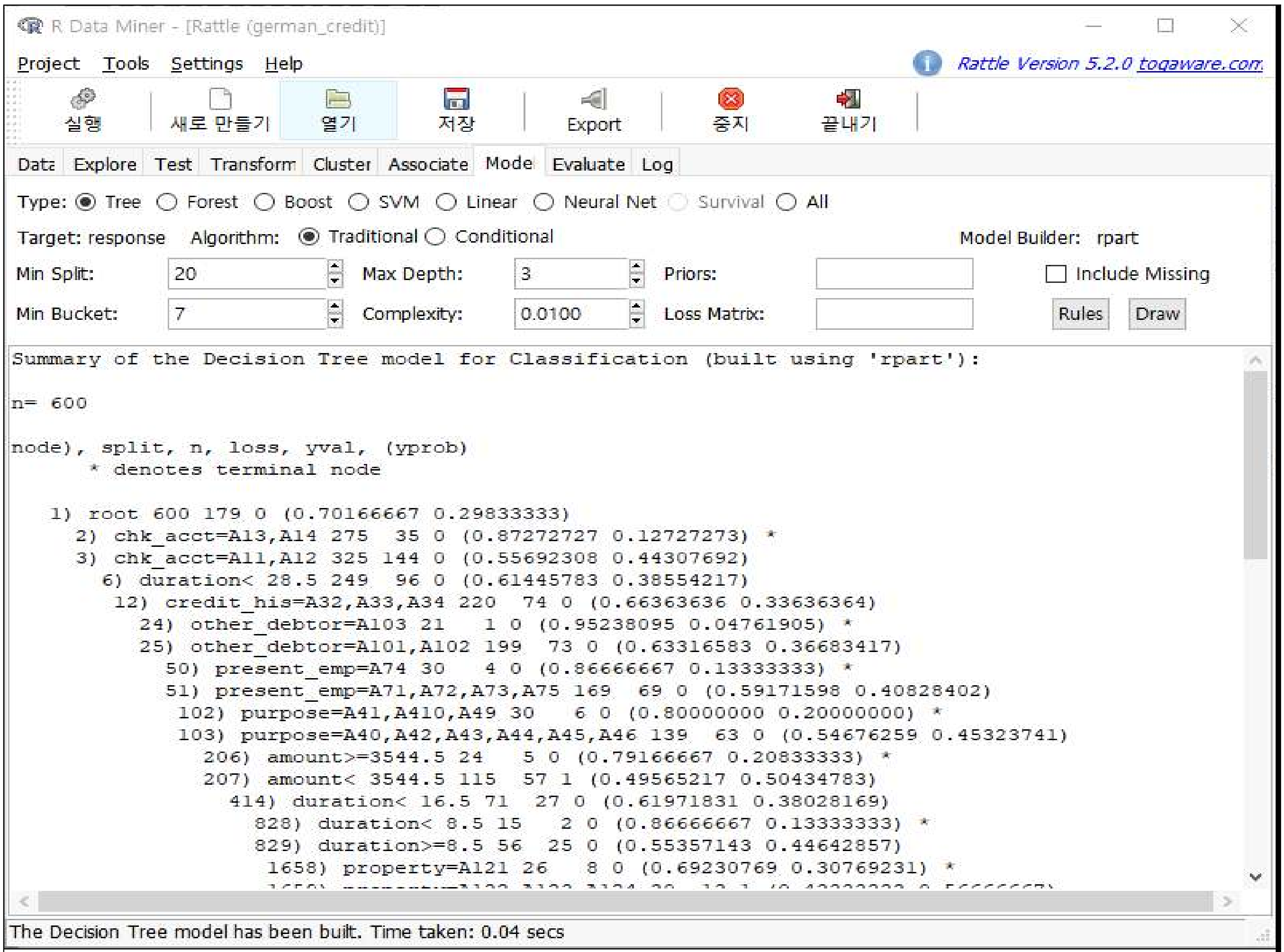
☑ Partition 60/20/20 Seed: 42 View Edit

● Input ● Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

| No. | Variable | Data Type | Input | Target | Risk | Ident | Ignore | Weight | Comment |
|---|---|---|---|---|---|---|---|---|---|
| 8 | installment_rate | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 4 |
| 9 | sex | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 4 |
| 10 | other_debtor | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 3 |
| 11 | present_resid | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 4 |
| 12 | property | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 4 |
| 13 | age | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 53 |
| 14 | other_install | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 3 |
| 15 | housing | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 3 |
| 16 | n_credits | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 4 |
| 17 | job | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 4 |
| 18 | n_people | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 2 |
| 19 | telephone | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 2 |
| 20 | foreign | Categoric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 2 |
| 21 | response | Categoric | ○ | ● | ○ | ○ | ○ | ○ | Unique: 2 |

Roles noted. 1,000 observations and 20 input variables. The target is response. Categoric 2. Classification models enabled.


R Data Miner - [Rattle (german_credit)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
실행 새로 만들기 열기 저장 Export 중지 끝내기
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rules Draw

Type: ● Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

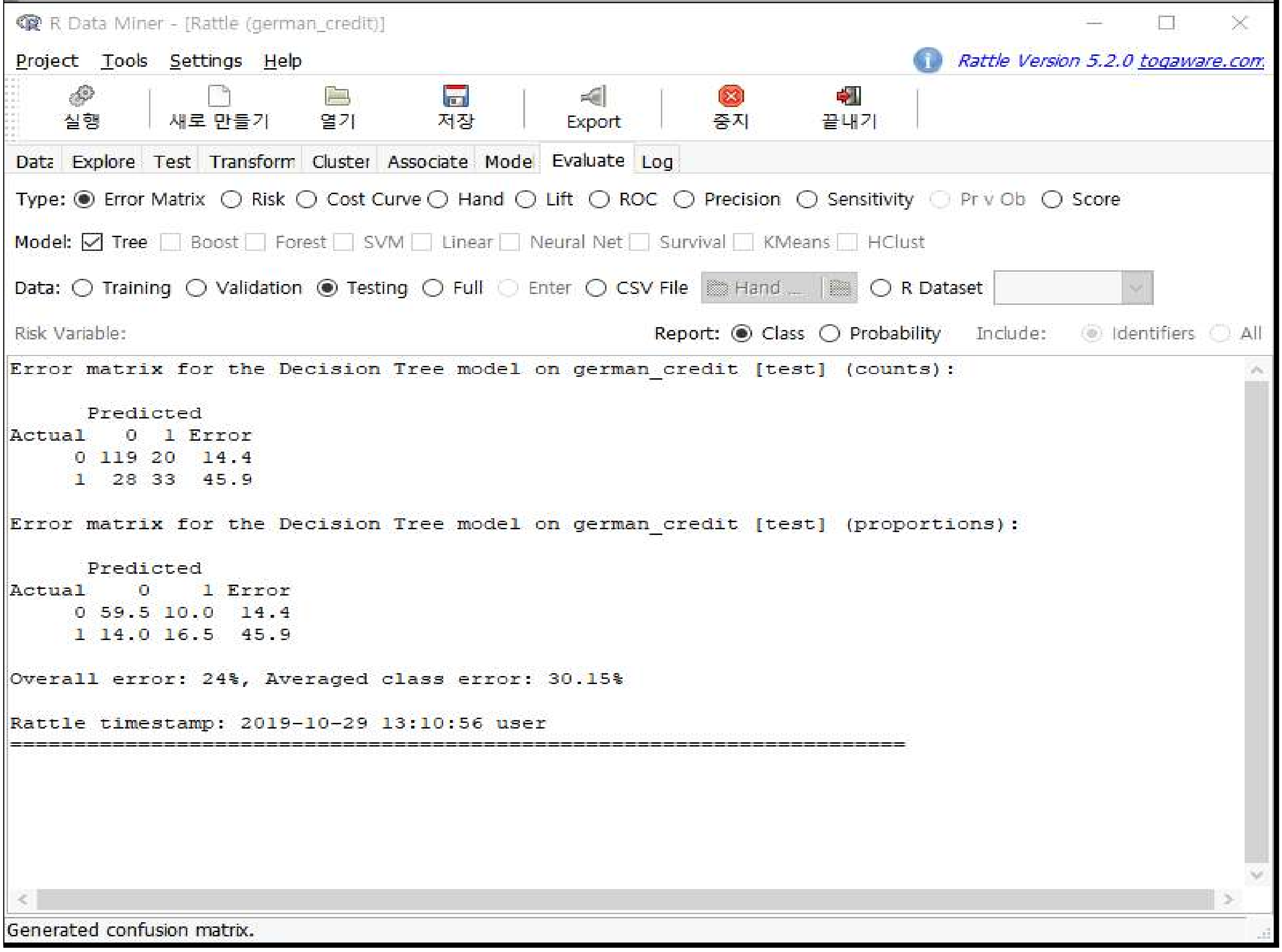
Target: response Algorithm: ● Traditional ○ Conditional Model Builder: rpart

Min Split: 20 Max Depth: 3 Priors: □ Include Missing

Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 600

node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 600 179 0 (0.70166667 0.29833333)
     2) chk_acct=A13,A14 275  35 0 (0.87272727 0.12727273) *
     3) chk_acct=A11,A12 325 144 0 (0.55692308 0.44307692)
       6) duration< 28.5 249  96 0 (0.61445783 0.38554217)
        12) credit_his=A32,A33,A34 220  74 0 (0.66363636 0.33636364)
          24) other_debtor=A103 21   1 0 (0.95238095 0.04761905) *
          25) other_debtor=A101,A102 199  73 0 (0.63316583 0.36683417)
            50) present_emp=A74 30   4 0 (0.86666667 0.13333333) *
            51) present_emp=A71,A72,A73,A75 169  69 0 (0.59171598 0.40828402)
             102) purpose=A41,A410,A49 30   6 0 (0.80000000 0.20000000) *
             103) purpose=A40,A42,A43,A44,A45,A46 139  63 0 (0.54676259 0.45323741)
               206) amount>=3544.5 24   5 0 (0.79166667 0.20833333) *
               207) amount< 3544.5 115  57 1 (0.49565217 0.50434783)
                 414) duration< 16.5 71  27 0 (0.61971831 0.38028169)
                   828) duration< 8.5 15   2 0 (0.86666667 0.13333333) *
                   829) duration>=8.5 56  25 0 (0.55357143 0.44642857)
                    1658) property=A121 26   8 0 (0.69230769 0.30769231) *
```

The Decision Tree model has been built. Time taken: 0.04 secs


R Data Miner - [Rattle (german_credit)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
실행 새로 만들기 열기 저장 Export 중지 끝내기
Data Explore Test Transform Cluster Associate Model Evaluate Log
Hand ...

Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

Model: ☑ Tree □ Boost □ Forest □ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust

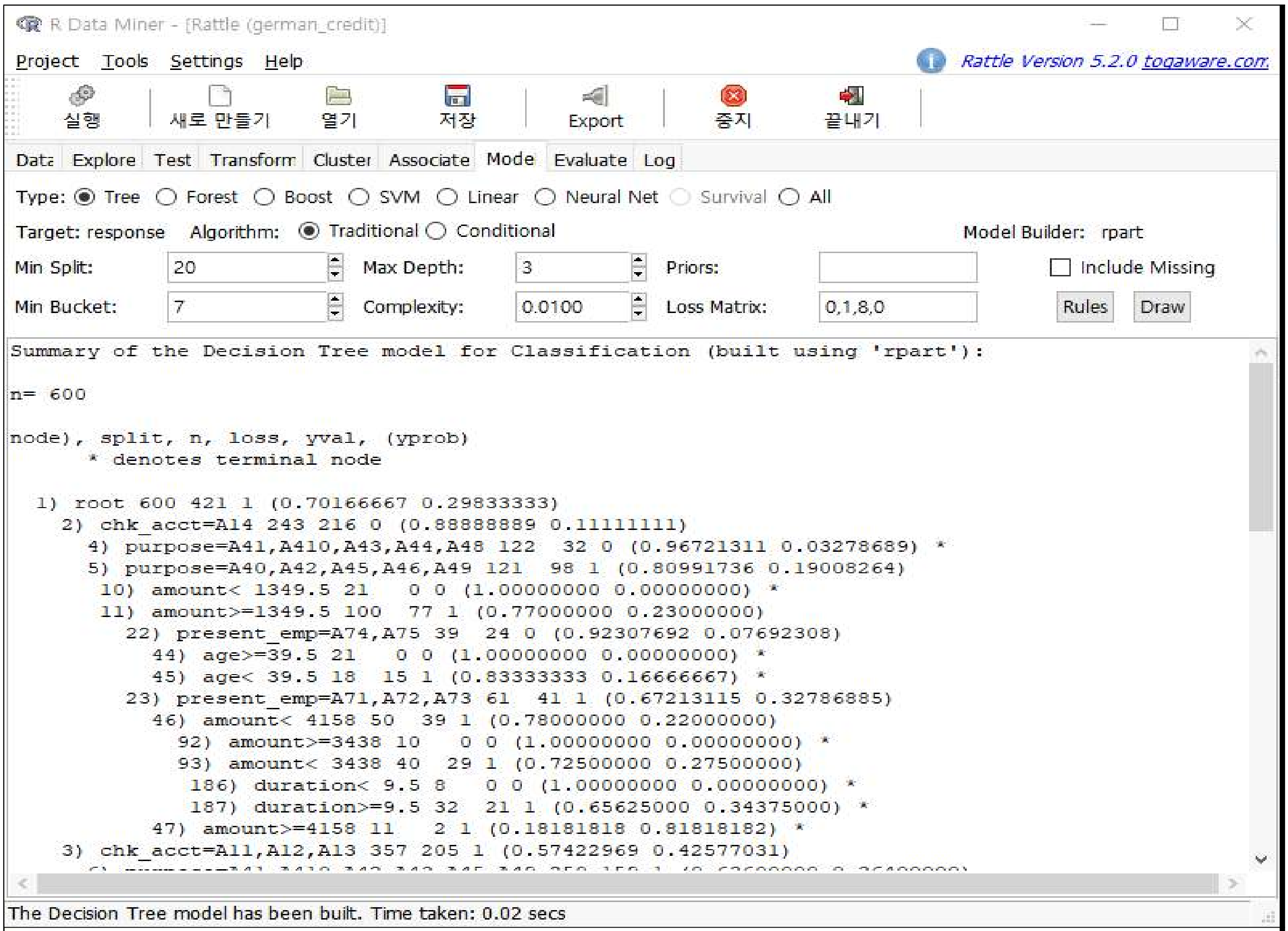
Data: ○ Training ○ Validation ● Testing ○ Full ○ Enter ○ CSV File ○ R Dataset

Risk Variable: Report: ● Class ○ Probability Include: ● Identifiers ○ All

```
Error matrix for the Decision Tree model on german_credit [test] (counts):

      Predicted
Actual   0  1 Error
     0 119 20  14.4
     1  28 33  45.9

Error matrix for the Decision Tree model on german_credit [test] (proportions):

      Predicted
Actual    0    1 Error
     0 59.5 10.0  14.4
     1 14.0 16.5  45.9

Overall error: 24%, Averaged class error: 30.15%

Rattle timestamp: 2019-10-29 13:10:56 user
======================================================================
```

Generated confusion matrix.

# Decision Tree vs Logistic Regression


R Data Miner - [Rattle (german_credit)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
실행 새로 만들기 열기 저장 Export 중지 끝내기
Data Explore Test Transform Cluster Associate Model Evaluate Log
Rules Draw

Type: ● Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

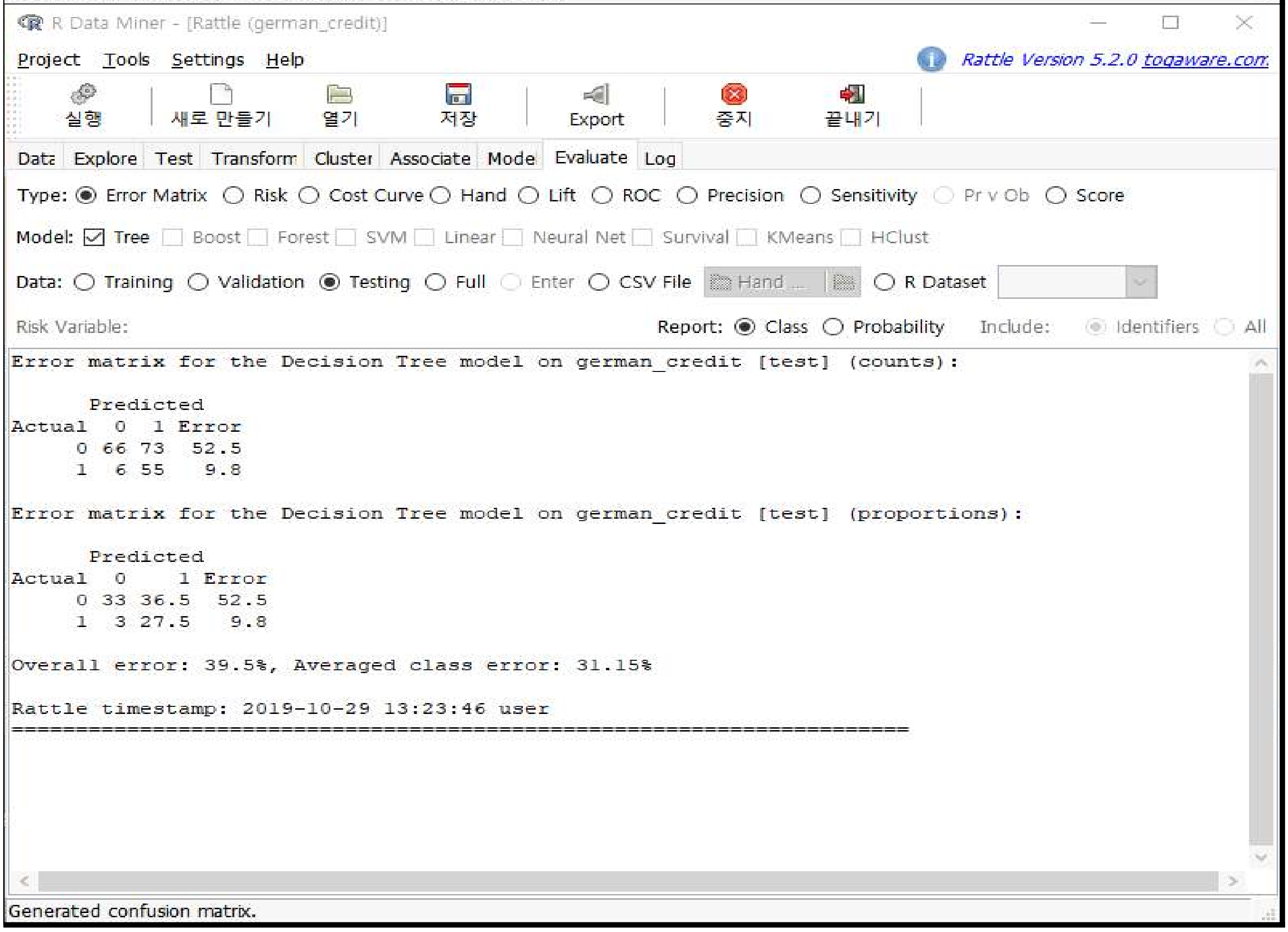
Target: response Algorithm: ● Traditional ○ Conditional Model Builder: rpart

Min Split: 20 Max Depth: 3 Priors: □ Include Missing

Min Bucket: 7 Complexity: 0.0100 Loss Matrix: 0,1,8,0

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 600

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 600 421 1 (0.70166667 0.29833333)
    2) chk_acct=A14 243 216 0 (0.88888889 0.11111111)
      4) purpose=A41,A410,A43,A44,A48 122  32 0 (0.96721311 0.03278689) *
      5) purpose=A40,A42,A45,A46,A49 121  98 1 (0.80991736 0.19008264)
       10) amount< 1349.5 21   0 0 (1.00000000 0.00000000) *
       11) amount>=1349.5 100  77 1 (0.77000000 0.23000000)
         22) present_emp=A74,A75 39  24 0 (0.92307692 0.07692308)
           44) age>=39.5 21   0 0 (1.00000000 0.00000000) *
           45) age< 39.5 18  15 1 (0.83333333 0.16666667) *
         23) present_emp=A71,A72,A73 61  41 1 (0.67213115 0.32786885)
           46) amount< 4158 50  39 1 (0.78000000 0.22000000)
             92) amount>=3438 10   0 0 (1.00000000 0.00000000) *
             93) amount< 3438 40  29 1 (0.72500000 0.27500000)
              186) duration< 9.5 8   0 0 (1.00000000 0.00000000) *
              187) duration>=9.5 32  21 1 (0.65625000 0.34375000) *
           47) amount>=4158 11   2 1 (0.18181818 0.81818182) *
    3) chk_acct=A11,A12,A13 357 205 1 (0.57422969 0.42577031)
```

The Decision Tree model has been built. Time taken: 0.02 secs


R Data Miner - [Rattle (german_credit)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
실행 새로 만들기 열기 저장 Export 중지 끝내기
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

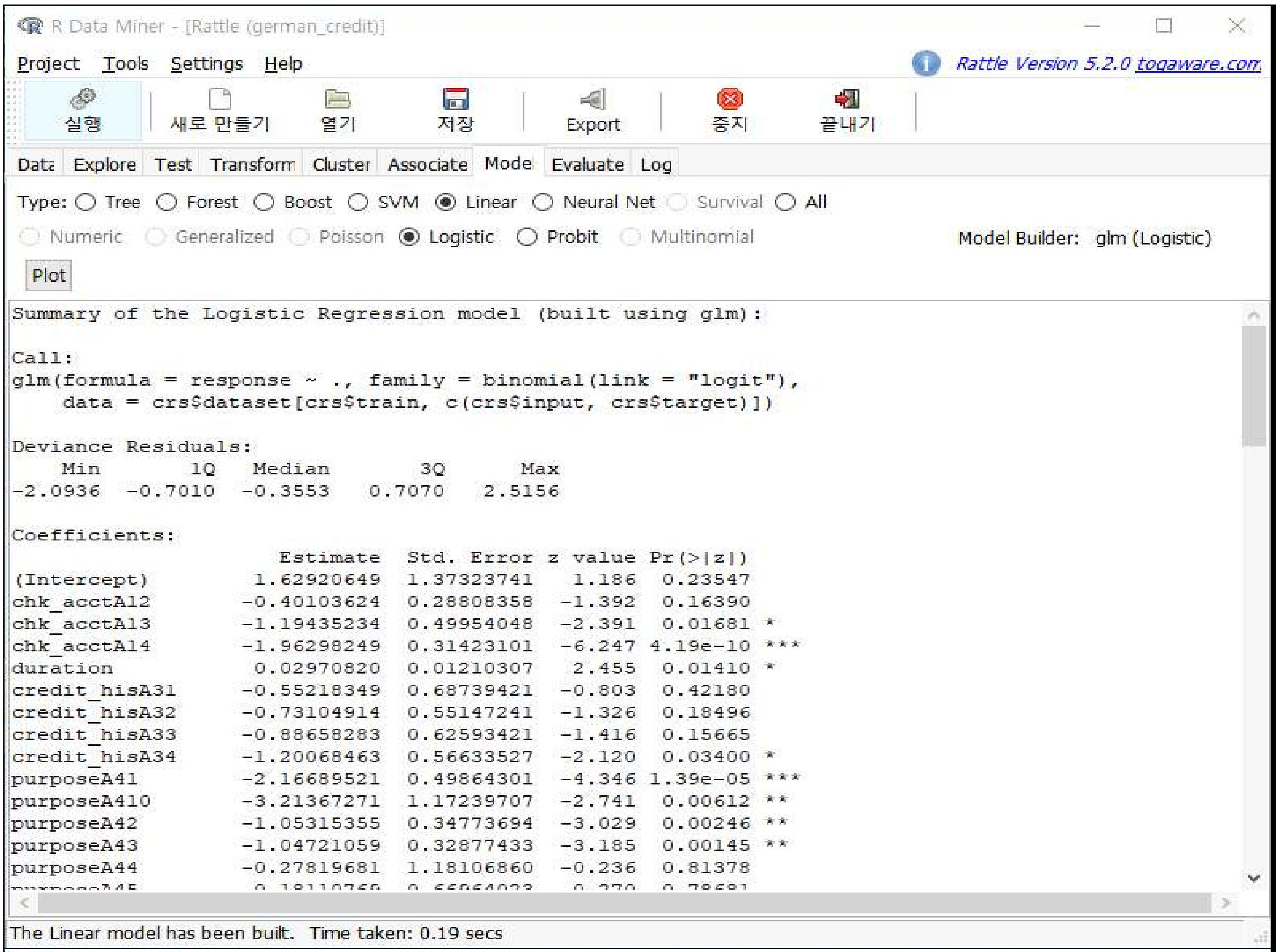
Model: ☑ Tree □ Boost □ Forest □ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ○ Training ○ Validation ● Testing ○ Full ○ Enter ○ CSV File Hand ... ○ R Dataset

Risk Variable: Report: ● Class ○ Probability Include: ● Identifiers ○ All

```
Error matrix for the Decision Tree model on german_credit [test] (counts):

      Predicted
Actual  0  1 Error
     0 66 73  52.5
     1  6 55   9.8

Error matrix for the Decision Tree model on german_credit [test] (proportions):

      Predicted
Actual  0    1 Error
     0 33 36.5  52.5
     1  3 27.5   9.8

Overall error: 39.5%, Averaged class error: 31.15%

Rattle timestamp: 2019-10-29 13:23:46 user
======================================================================
```

Generated confusion matrix.


R Data Miner - [Rattle (german_credit)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
실행
새로 만들기
열기
저장
Export
중지
끝내기
Data Explore Test Transform Cluster Associate Model Evaluate Log
Plot

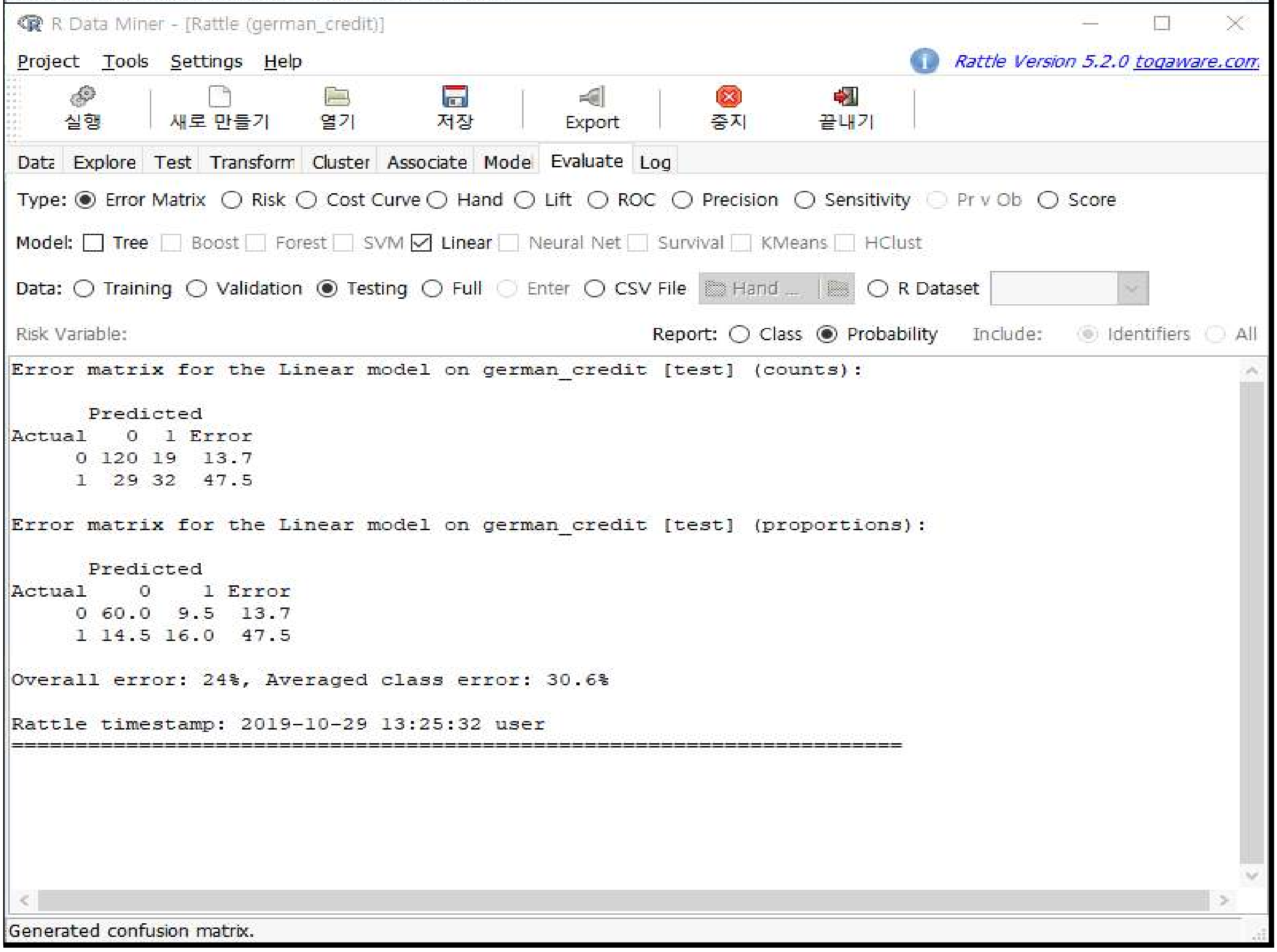
Type: ○ Tree ○ Forest ○ Boost ○ SVM ● Linear ○ Neural Net ○ Survival ○ All

○ Numeric ○ Generalized ○ Poisson ● Logistic ○ Probit ○ Multinomial

Model Builder: glm (Logistic)

```
Summary of the Logistic Regression model (built using glm):

Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = crs$dataset[crs$train, c(crs$input, crs$target)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0936  -0.7010  -0.3553   0.7070   2.5156

Coefficients:
                    Estimate  Std. Error z value Pr(>|z|)
(Intercept)       1.62920649  1.37323741   1.186  0.23547
chk_acctA12      -0.40103624  0.28808358  -1.392  0.16390
chk_acctA13      -1.19435234  0.49954048  -2.391  0.01681 *
chk_acctA14      -1.96298249  0.31423101  -6.247 4.19e-10 ***
duration          0.02970820  0.01210307   2.455  0.01410 *
credit_hisA31    -0.55218349  0.68739421  -0.803  0.42180
credit_hisA32    -0.73104914  0.55147241  -1.326  0.18496
credit_hisA33    -0.88658283  0.62593421  -1.416  0.15665
credit_hisA34    -1.20068463  0.56633527  -2.120  0.03400 *
purposeA41       -2.16689521  0.49864301  -4.346 1.39e-05 ***
purposeA410      -3.21367271  1.17239707  -2.741  0.00612 **
purposeA42       -1.05315355  0.34773694  -3.029  0.00246 **
purposeA43       -1.04721059  0.32877433  -3.185  0.00145 **
purposeA44       -0.27819681  1.18106860  -0.236  0.81378
purposeA45        0.18110760  0.66964023   0.270  0.78681
```

The Linear model has been built.  Time taken: 0.19 secs


R Data Miner - [Rattle (german_credit)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
실행
새로 만들기
열기
저장
Export
중지
끝내기
Data Explore Test Transform Cluster Associate Model Evaluate Log
Hand ...

Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

Model: □ Tree □ Boost □ Forest □ SVM ☑ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ○ Training ○ Validation ● Testing ○ Full ○ Enter ○ CSV File ○ R Dataset

Risk Variable:

Report: ○ Class ● Probability Include: ● Identifiers ○ All

```
Error matrix for the Linear model on german_credit [test] (counts):

      Predicted
Actual   0  1 Error
     0 120 19  13.7
     1  29 32  47.5

Error matrix for the Linear model on german_credit [test] (proportions):

      Predicted
Actual    0    1 Error
     0 60.0  9.5  13.7
     1 14.5 16.0  47.5

Overall error: 24%, Averaged class error: 30.6%

Rattle timestamp: 2019-10-29 13:25:32 user
======================================================================
```

Generated confusion matrix.